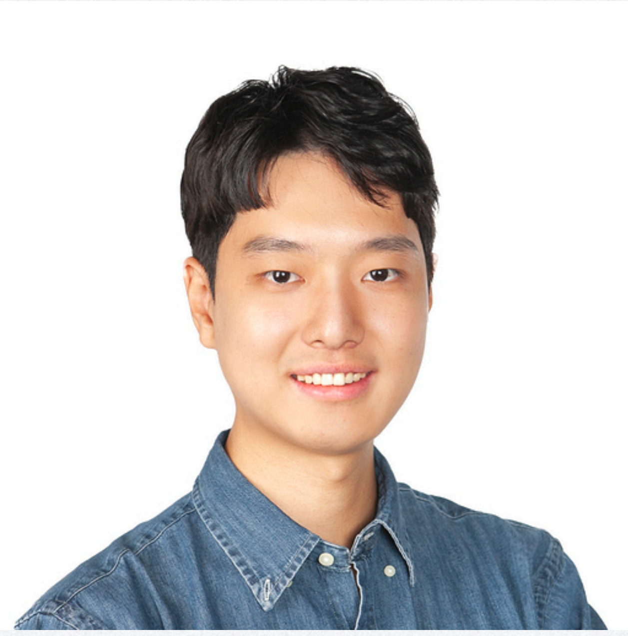
The Linear Representation Hypothesis and the Geometry of Large Language Models

NeurIPS 2023 Workshop on Causal Representation Learning
December 15, 2023, New Orleans, Louisiana, USA


Kiho Park
Stat @ UChicago

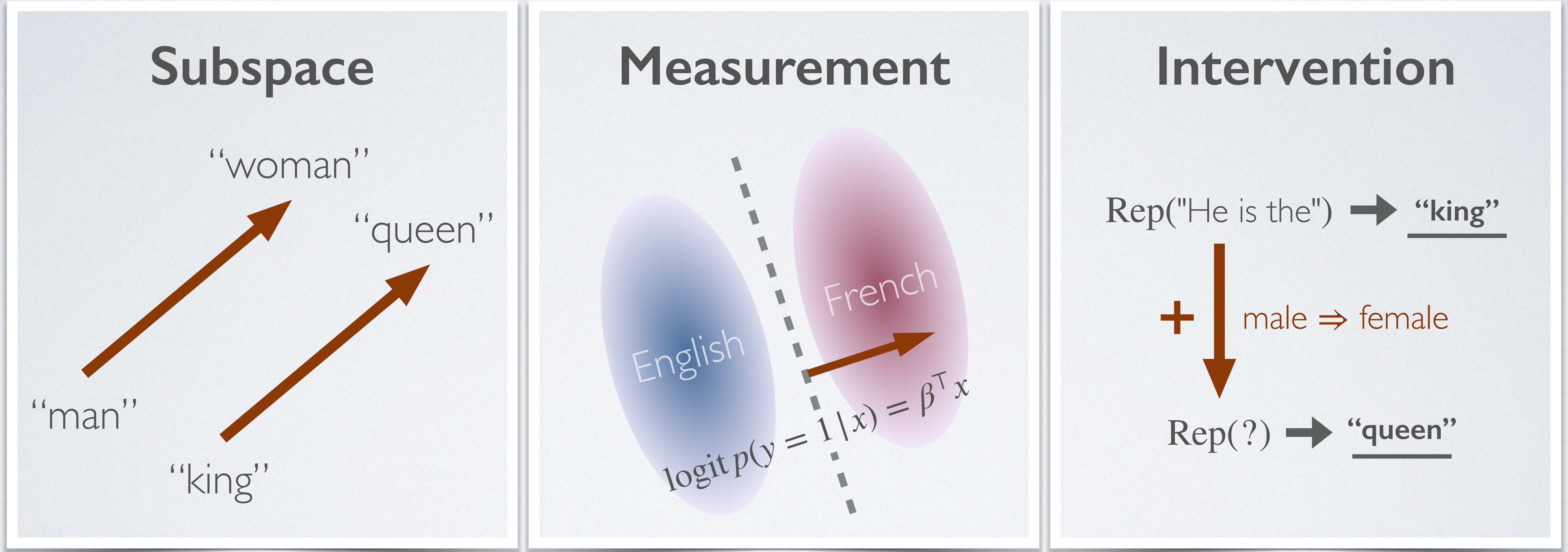
Yo Joong (YJ) Choe
DSI @ UChicago

Victor Veitch
Stat & DSI @ UChicago

Linear Representation Hypothesis

High-level concepts are represented ***linearly***
as directions in the representation space

What Does “Linear” Even Mean?



Problem: not clear how these relate, nor which is the ‘right’ notion

Strategy

- Define what "concept" means
- Formalize subspace notion of linearity
- Use softmax structure to connect to measurement and intervention

Background on LLMs

Context

"He is the" $X \longrightarrow Y$

Next word

"king"

"man"

"PhD"

$\vdots$

Embedding

$\lambda(x) \in \mathbb{R}^d$

Softmax

$\mathbb{P}(y \mid x) \propto \exp(\lambda(x)^\top \gamma(y))$

Unembedding

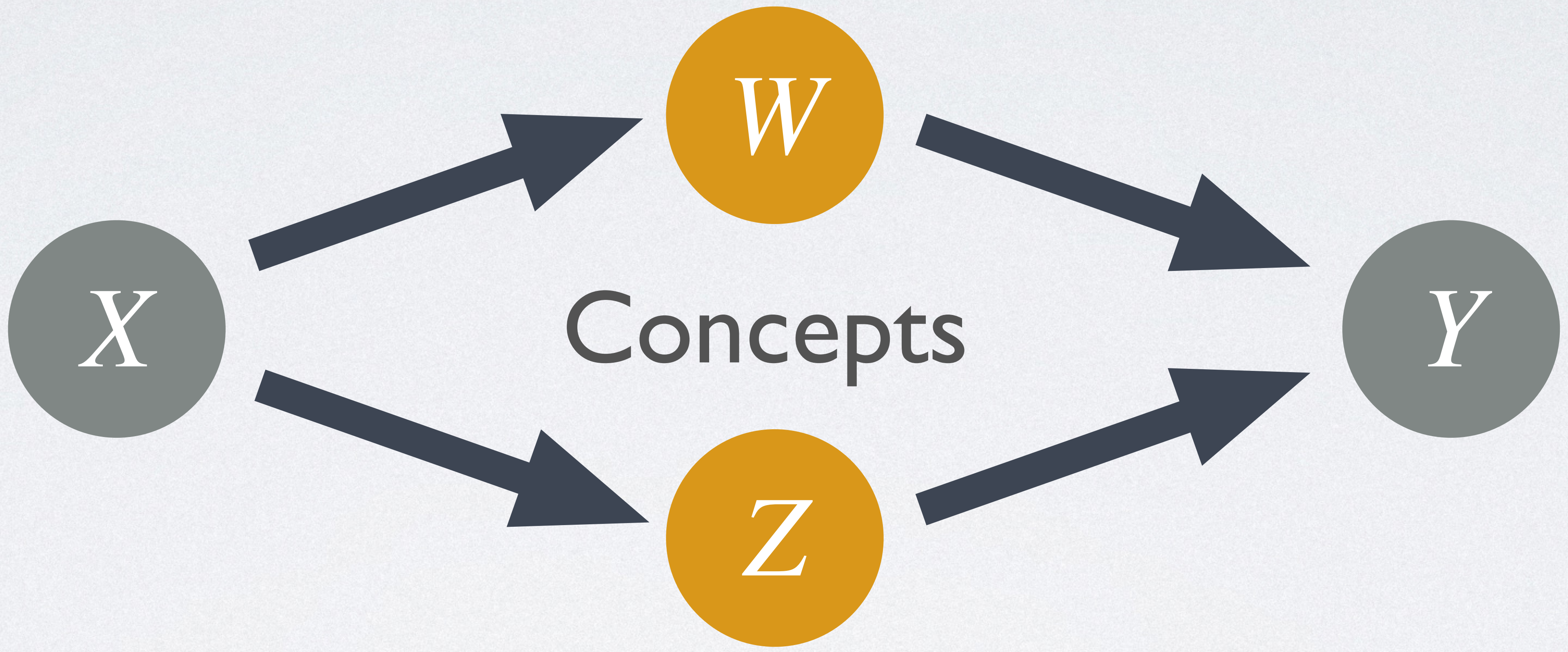
$\gamma(y) \in \mathbb{R}^d$

Concepts in LLMs

male $(W = 0) \Rightarrow$ female $(W = 1)$

Context

"He is the"



Next word

"king"	$Y(0,0)$
"queen"	$Y(1,0)$
"roi"	$Y(0,1)$
"reine"	$Y(1,1)$

English $(Z = 0) \Rightarrow$ French $(Z = 1)$

A concept W is defined by a set of counterfactual outputs $Y(W = 0), Y(W = 1)$

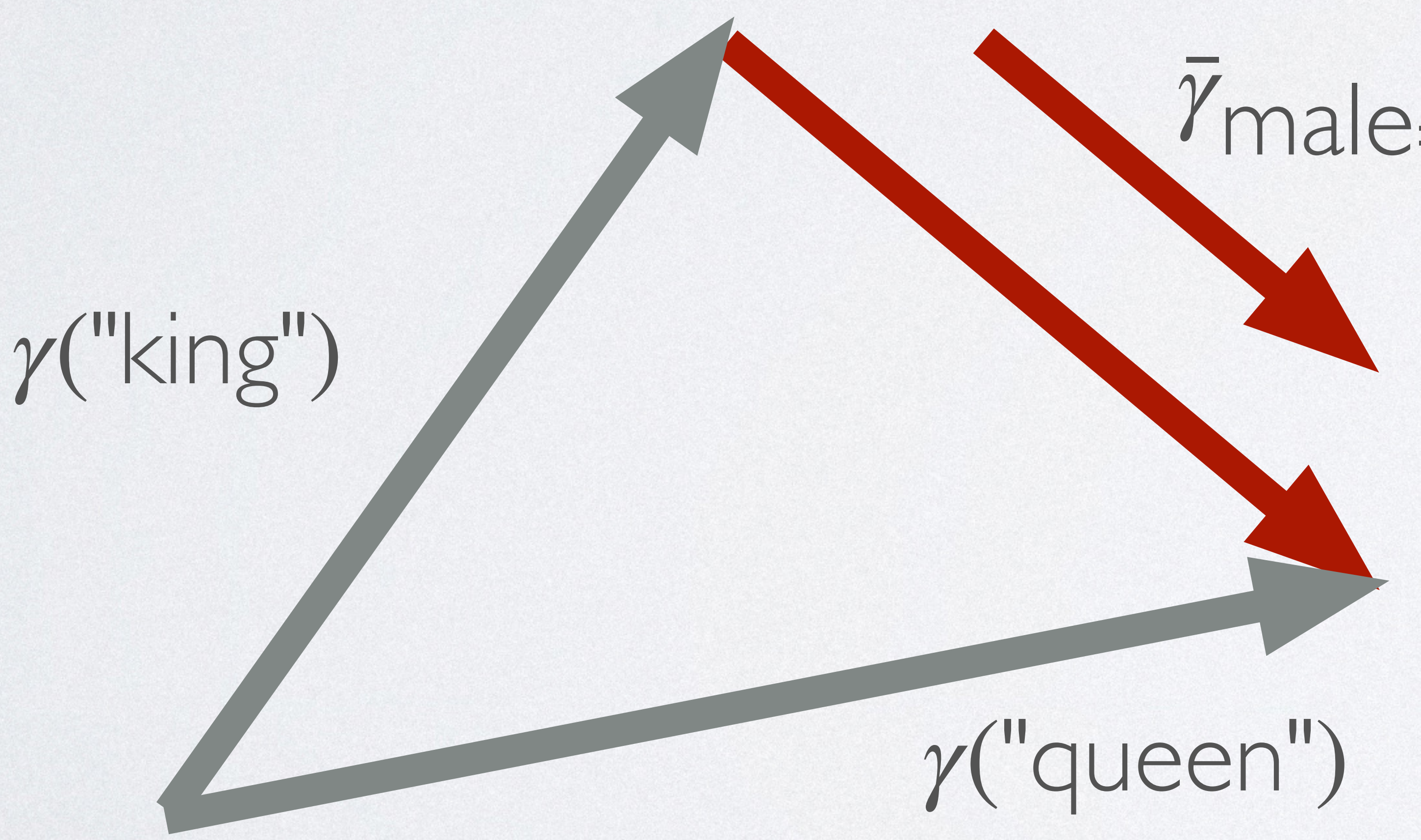
Concepts W, Z are *causally separable* if $Y(w, z)$ is well-defined

Subspace Notions of Linear Representations

Unembedding Representation $\bar{\gamma}_W$

$$\gamma(Y(1)) - \gamma(Y(0)) = \alpha\bar{\gamma}_W \quad (\alpha > 0)$$

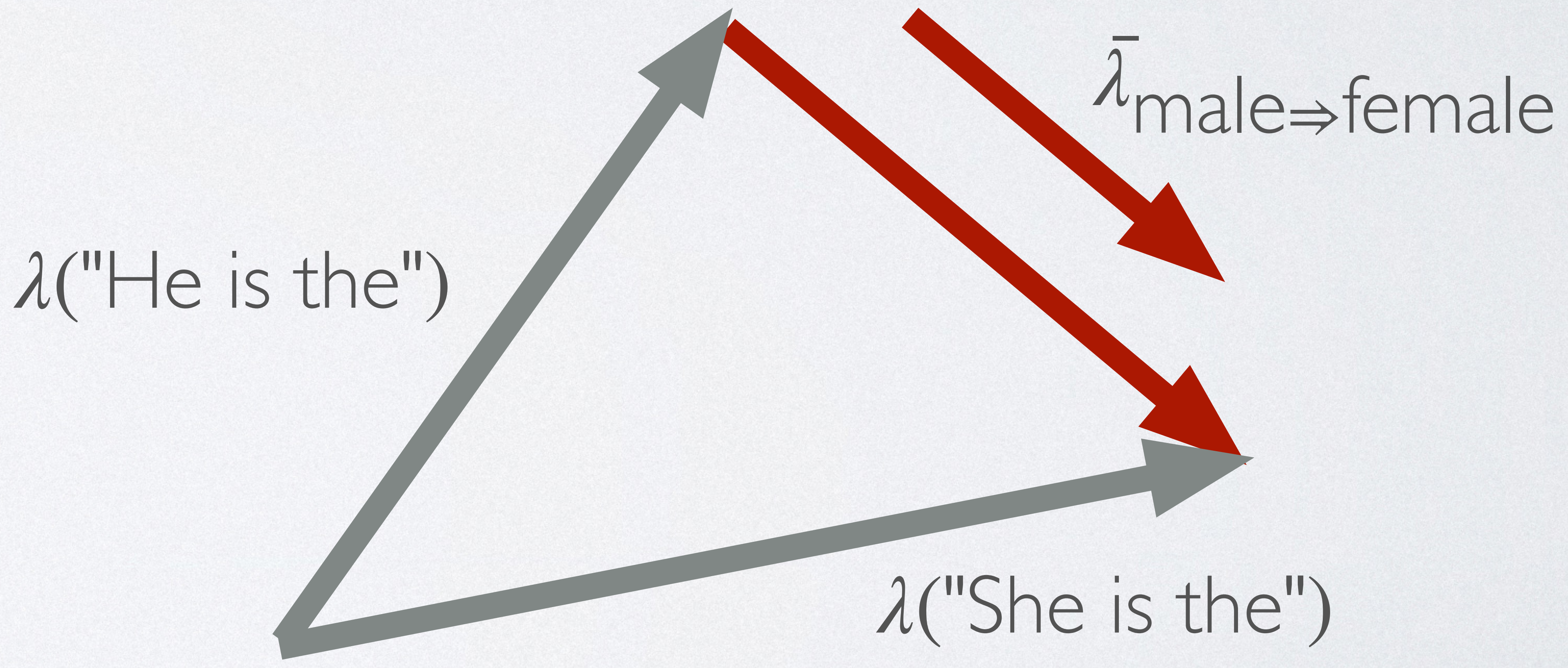
for any counterfactual pairs $(Y(0), Y(1))$



Embedding Representation $\bar{\lambda}_W$

$$\lambda_1 - \lambda_0 = \alpha'\bar{\lambda}_W \quad (\alpha' > 0)$$

for any counterfactual pairs (λ_0, λ_1)



Connection to Measurement and Intervention

Unembedding Representation $\bar{\gamma}_W$

$$\gamma(Y(1)) - \gamma(Y(0)) = \alpha\bar{\gamma}_W \quad (\alpha > 0)$$

for any counterfactual pairs $(Y(0), Y(1))$

Theorem

Measurement Representation

$$\operatorname{logit} \mathbb{P}(W = 1 \mid \lambda) = \alpha\lambda^\top\bar{\gamma}_W$$

Embedding Representation $\bar{\lambda}_W$

$$\lambda_1 - \lambda_0 = \alpha'\bar{\lambda}_W \quad (\alpha' > 0)$$

for any counterfactual pairs (λ_0, λ_1)

Theorem

Intervention Representation

$\mathbb{P}(W = 1 \mid Z, \lambda + c\bar{\lambda}_W)$ increasing in $c \in \mathbb{R}$

$\mathbb{P}(Z = 1 \mid W, \lambda + c\bar{\lambda}_W)$ constant in $c \in \mathbb{R}$

Problems

- How do the unembedding and embedding representations relate?

- What is the right inner product for the representation space?

$$\mathbb{P}(y \mid x) \propto \exp(\lambda(x)^\top \gamma(y)) = \exp\left(\left(A^{-\top} \lambda(x) \right)^\top \left(A \gamma(y) \right) \right)$$

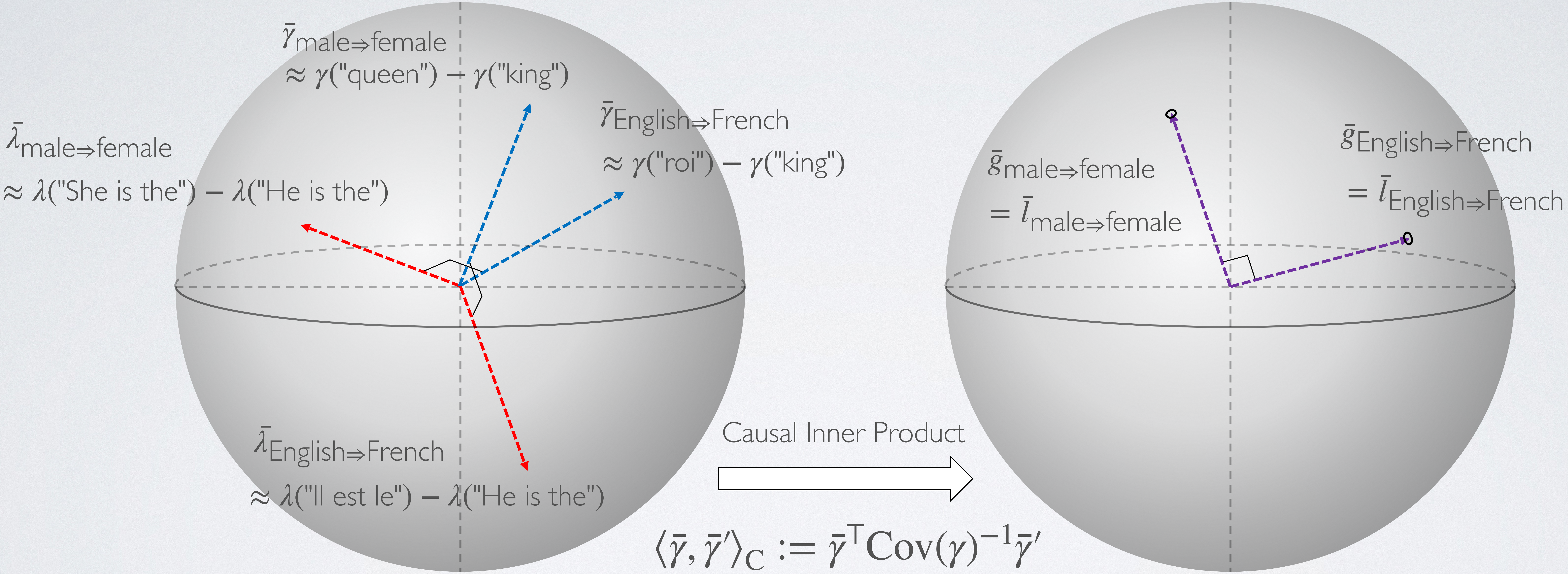
for any invertible $A \in \mathbb{R}^{d \times d}$

Causal Inner Product

- **Definition:** $\langle \cdot , \cdot \rangle_{\mathrm{C}}$ is a *causal inner product* if $\langle \bar{\gamma}_W, \bar{\gamma}_Z \rangle_{\mathrm{C}} = 0$ whenever W and Z are causally separable

- **Theorem:** The causal inner product unifies the unembedding and embedding representations via $\langle \bar{\gamma}_W, \cdot \rangle_{\mathrm{C}} = (\bar{\lambda}_W)^\top$ (by Riesz isomorphism)

Causal Inner Product



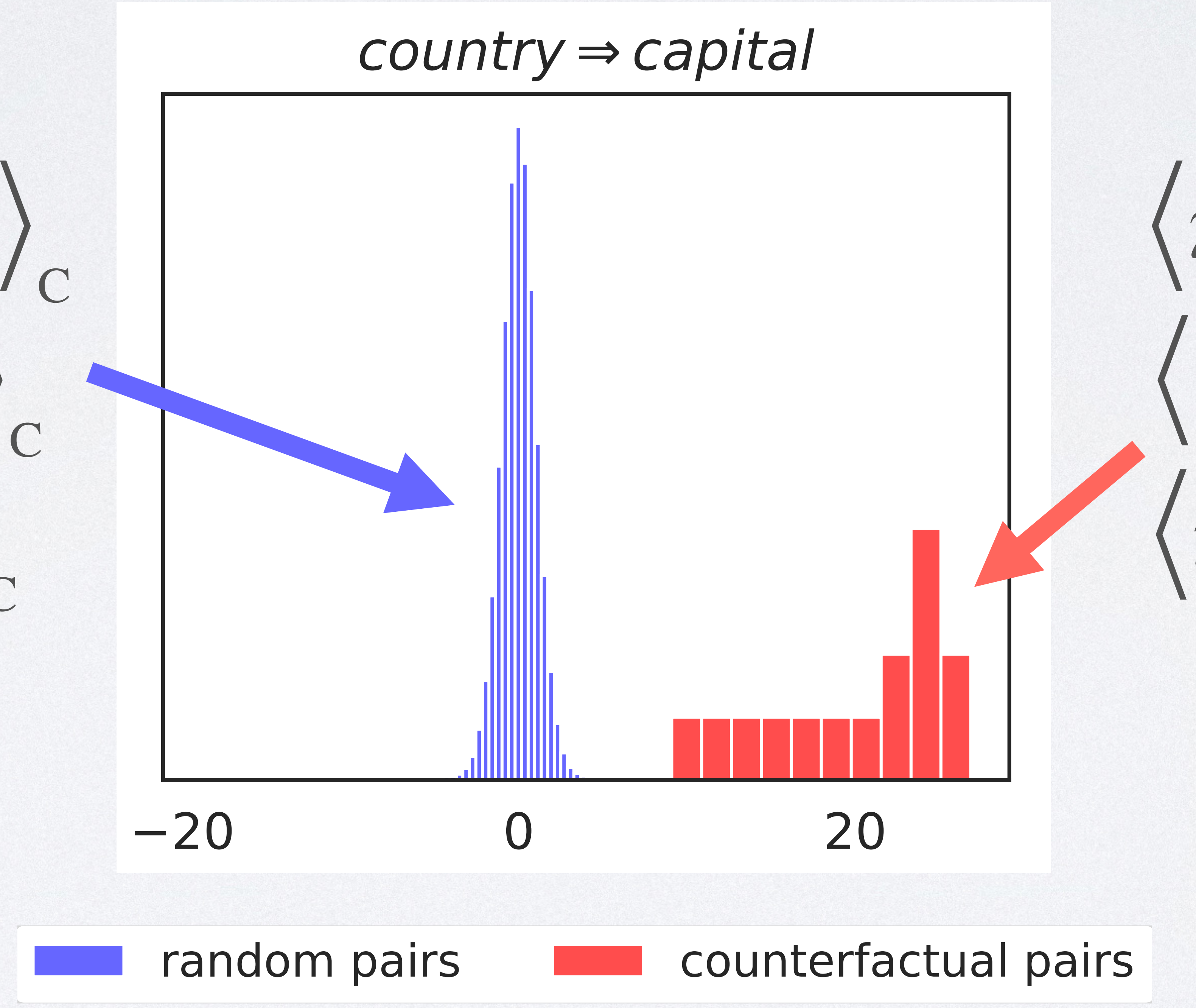
$$\langle \bar{\gamma}, \bar{\gamma}' \rangle_{\mathrm{C}} := \bar{\gamma}^{\top} \mathrm{Cov}(\gamma)^{-1} \bar{\gamma}'$$

Experiments

Linear Representations Exist

$$\bar{\gamma}_W := \frac{1}{n_W} \sum_{i=1}^{n_W} \gamma(y_i(1)) - \gamma(y_i(0))$$

$$\left\langle \gamma(\text{"volcano"}) - \gamma(\text{"wrote"}), \bar{\gamma}_{\text{country}\Rightarrow\text{capital}} \right\rangle_{\mathrm{C}}$$
$$\left\langle \gamma(\text{"chairs"}) - \gamma(\text{"happy"}), \bar{\gamma}_{\text{country}\Rightarrow\text{capital}} \right\rangle_{\mathrm{C}}$$
$$\left\langle \gamma(\text{"April"}) - \gamma(\text{"jump"}), \bar{\gamma}_{\text{country}\Rightarrow\text{capital}} \right\rangle_{\mathrm{C}}$$
$$\vdots$$

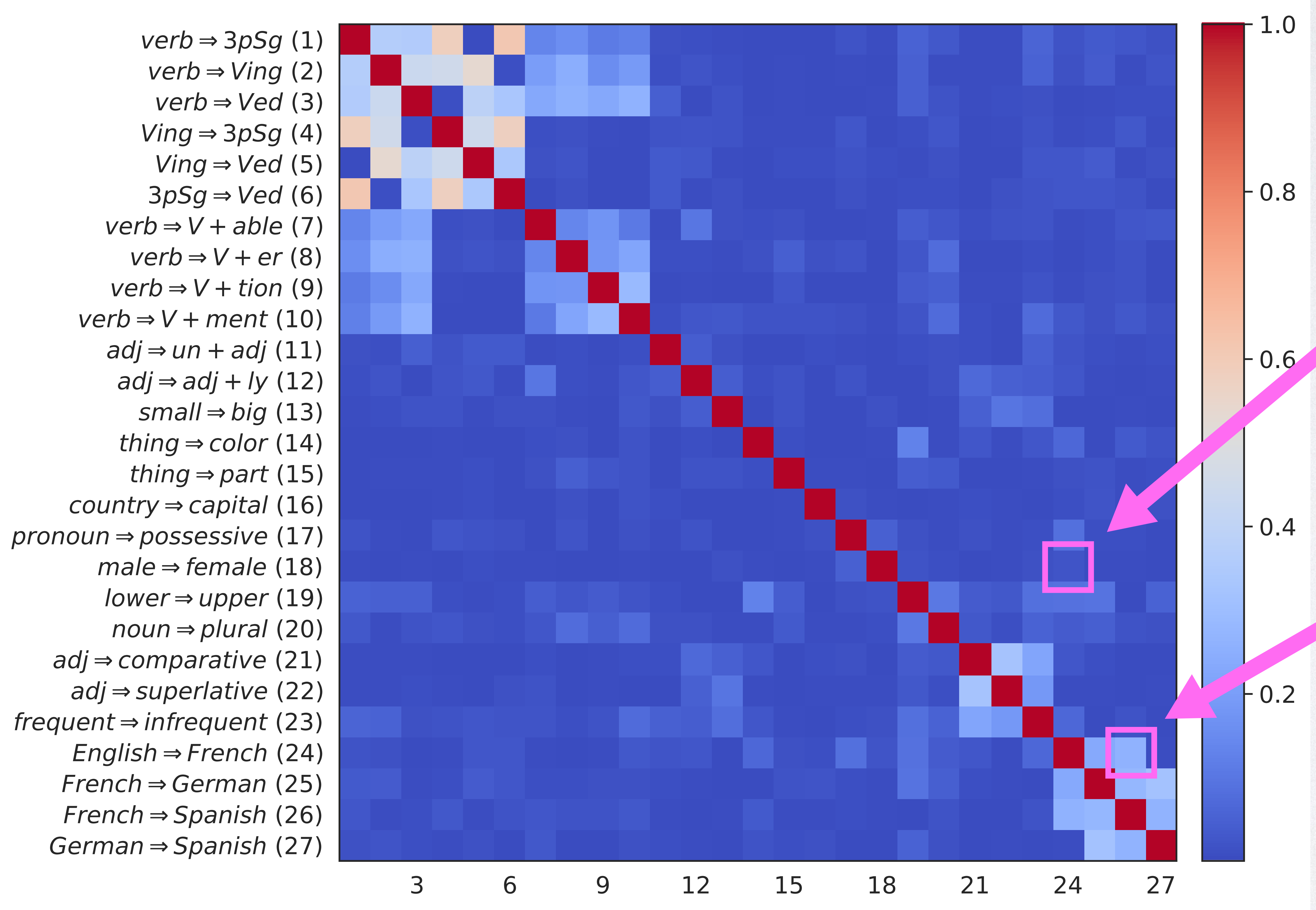


$$\left\langle \gamma(\text{"Paris"}) - \gamma(\text{"France"}), \bar{\gamma}_{\text{country}\Rightarrow\text{capital}} \right\rangle_{\mathrm{C}}$$
$$\left\langle \gamma(\text{"Tokyo"}) - \gamma(\text{"Japan"}), \bar{\gamma}_{\text{country}\Rightarrow\text{capital}} \right\rangle_{\mathrm{C}}$$
$$\left\langle \gamma(\text{"Cairo"}) - \gamma(\text{"Egypt"}), \bar{\gamma}_{\text{country}\Rightarrow\text{capital}} \right\rangle_{\mathrm{C}}$$
$$\vdots$$

Experiment with LLaMA-2-7B model

Causally Separable Concepts Are Represented Orthogonally Under the Causal Inner Product

$$|\langle \bar{\gamma}_W, \bar{\gamma}_{W'} \rangle_C| = |\bar{\gamma}_W^\top \mathrm{Cov}(\gamma)^{-1} \bar{\gamma}_{W'}|$$



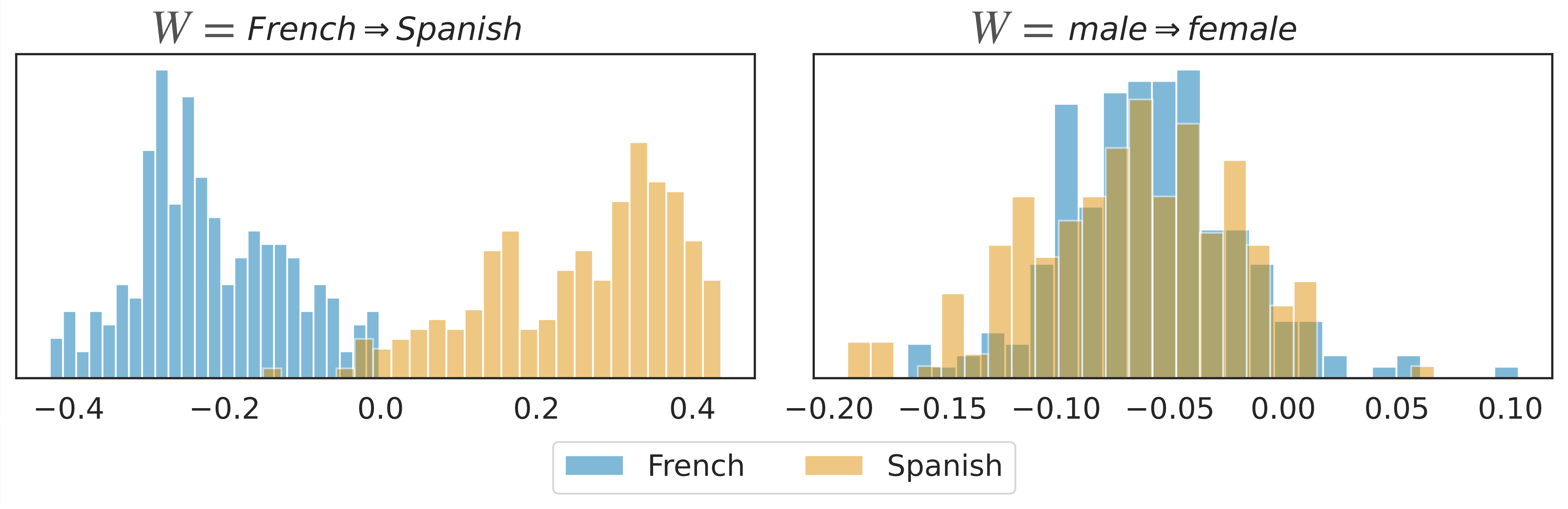
$$\left| \left\langle \bar{\gamma}_{\text{male}\Rightarrow\text{female}}, \bar{\gamma}_{\text{English}\Rightarrow\text{French}} \right\rangle_C \right| \approx 0$$

$$\left| \left\langle \bar{\gamma}_{\text{English}\Rightarrow\text{French}}, \bar{\gamma}_{\text{French}\Rightarrow\text{Spanish}} \right\rangle_C \right| \gg 0$$

Experiment with LLaMA-2-7B model

Unembedding Representation Yields Linear Probe

$$\bar{\gamma}_W^\top \lambda(x^{\mathrm{fr}}) \text{ vs. } \bar{\gamma}_W^\top \lambda(x^{\mathrm{es}})$$

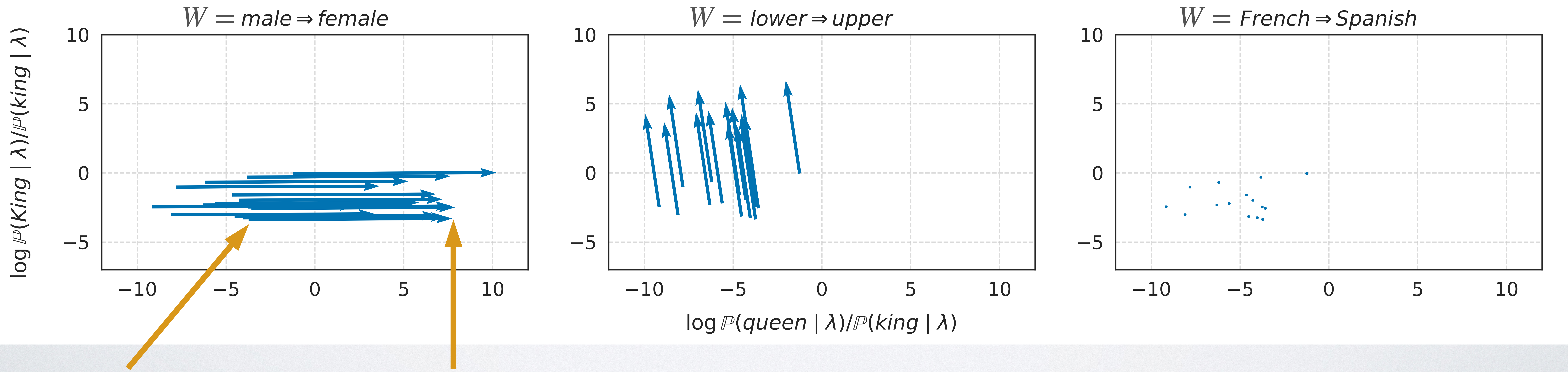


Experiment with LLaMA-2-7B model

Embedding Representation Yields Steering Vector

$$\bar{\lambda}_W := \mathrm{Cov}^{-1}(\gamma)\bar{\gamma}_W$$

Changes in logits resulting from adding $\alpha\bar{\lambda}_W$ to context embeddings



$\lambda = \lambda(\text{"Long live the"})$ $\lambda = \lambda(\text{"Long live the"}) + \alpha\bar{\lambda}_{\text{male}\Rightarrow\text{female}}$

Experiment with LLaMA-2-7B model

Summary

- We formalize several notions of 'linear representation' and show their interrelations
- With the right choice of inner product, every notion of linear representation unifies